Supporting Information

Ganguly et al. 10.1073/pnas.0904495106

SI Text

Statistical Methods for Significance and Uncertainty Estimates. The assignment of confidence bounds to the model-based climate projections makes several underlying assumptions about the nature of the data, specifically, for the differences between reanalysis (''observations'') and climate model simulations:

The application of a *t*-test and the use of sigma levels to assign confidence bounds implicitly assume a normal (Gaussian) distribution. This assumption needs to be tested, especially over the time period 2000–2007, which in turn is used to compute the bias and standard deviation.

The computation of bias assumes that the mean of the normal distribution is stationary over time, while the computation of uncertainty assumes the same is true for variance (i.e., heteroskedasticity is ruled out). The stationary assumptions in the differences cannot be tested for projections, since reanalysis (''observed'') data are not available for the future. However, as a proxy, we can test whether the stationary assumptions hold for previous time steps when climate model hindcasts can indeed be compared with reanalysis data.

Bias and variance are assumed to remain constant with projection lead times. This assumption is difficult to test, since reanalysis data do not exist for the future and because climate models evolve over time. The latter implies that accurate estimates of the bias and variance of the most recent models as a function of lead times typically cannot be obtained from prior versions of climate models. However, normal distributions in decadal differences, a lack of trend in the 2000–2007 bias and variance, as well as nonstationary bias and variance in hindcasts, suggest that these assumptions may not be unreasonable. However, we conjecture that the variance will increase with lead times.

First, we test the assumption that the daily differences between the reanalysis data and the CCSM 3.0 model projections are in fact normally distributed, especially for 2000–2007. The correspondence of the histogram with the fitted normal distribution is visually examined. In addition, we produce the quantile-quantile (Q-Q) plots, which test for normality. As shown in the top left panel of Fig. S1, the histogram of the differences does indeed follow a Gaussian (with mean −0.018) visually, which is reflected in the Q-Q plot as well. The assumptions of normality are further tested by comparing hindcasts from climate model simulations with reanalysis data for each decade from 1950 onward. As shown in Fig. S1, the decadal differences follow the Gaussian quite well, both visually from the histograms as well as from the Q-Q plots.

The mean or standard deviation of the normal distributions in Fig. S1 does not appear to change significantly over six consecutive decades. We test the stationary assumptions further in Fig. S2. Trends in the mean and variance of the differences, along with the confidence bounds along the trend lines, are examined through regression analysis. The top, middle, and bottom panels (there are two plots per panel) show the trends in the mean and variance on the daily difference data after the application of an annual moving average, a decadal moving average, and a 25-year moving average. At 95% confidence levels, the trend lines do not appear to deviate much from the zero-slope regression lines in a statistically significant manner, even though slight trends may exist for the 25-year moving averages. The annual moving averages of the differences do not exhibit any linear trends whatsoever in the mean or the variance, but there are interannual features, which only rarely rise above the 95% confidence bands. The decadal moving averages appear to exhibit a slight upward trend in the mean and downward trend in the variance, which appears slightly enhanced in the 25-year moving averages. However, the deviations from a zero-slope line are relatively small. While decadal and multidecadal features are present in the differences, these are typically small compared to the 95% confidence bands. Thus, we can derive a few broad conclusions from Fig. S2. First, there are no major discernable trends in the mean or variance of the differences that fall beyond the 95% confidence bounds of the regression line. Second, the regression line is not significantly different from a corresponding zero-slope line, which implies stationary mean and variance. However, a slight trend (higher for mean and lower for variance) in the successive moving averages is noted. Third, weak interannual, decadal, and multidecadal features are noted in the differences. While not strong enough to overwhelm the bounds of the linear regression, these features may have physical interpretations. Thus, their relations with the interannual, decadal, and multidecadal sea surface temperature anomalies (SSTA) may need to be investigated, especially since climate models are known to be weak in modeling SSTA and their impacts on regional climate.

Fig. 1 shows global average projections of temperatures with uncertainty bounds. Here we present the complete *t*-test results between the three scenarios and indicate which pairings show separation at 95% confidence ($\alpha = 0.05$). As shown in Table S1, both the A1FI and A2 scenarios are significantly higher than B1 from 2040 onward, while A1FI does not separate from A2 until 2060 and then converges again by 2100.

Uncertainty in climate projections have typically been computed based on multimodel (12) and initial-condition ensembles (15), as well as by comparing model hindcasts with observations (13). Multimodel ensembles attempt to capture the uncertainty caused by our incomplete understanding of the physics and assume that a statistical averaging of the results from multiple models will be better than any one alone. However, since any physical understanding developed by the community is typically embedded within all models and the unknown physics are not captured in any model, the uncertainties may be underestimated from the multimodel approach. As suggested in Fig. 1 (bottom left panel, 2000–2007), the variability in the observations are rarely completely captured by the smoothed model outputs. Initial-condition ensembles capture only one, and perhaps very minor, aspect of the overall uncertainty. Fig. 1 visually depicts that bounds generated by the ensemble ranges for A2 and B1 are relatively small (compared to the bounds developed by comparing with the observations) and that these bounds do not fluctuate significantly over time, which indicates relative insensitivity to initial conditions for these specific projections at global-average scales. Furthermore, uncertainty bounds generated from comparisons with observations and shown as three standard deviation levels around the bias-corrected mean are much larger in all cases compared to the bounds generated from the ensembles based on the median value for the central tendency at any given point and the minimum and maximum as the bounds. We confirmed that the standard deviations are different and statistically distinguishable based on the F-test (see Table S2), especially since the observation-based bounds are always larger than the initial-condition ensemble-based bounds.

Uncertainties based on comparing model hindcasts with observations can provide important insights on the predictability and systematic errors of climate models. However, given that climate models continually improve based on incoming infor-

mation, the hindcasts may be closer to observations than future projections; hence the uncertainty assessments in projections may be underestimated. On the other hand, results from prior model versions originally run in projection mode may be compared with current observations. However, since models improve continually, such an approach may overestimate the current uncertainty. Here we assume that the IPCC SRES scenario definitions (at least conceptually) and the CCSM 3.0 model have not changed significantly since 2000. Thus, here we have been able to compare the 2000–2007 model results almost in a forecast (rather than a hindcast) mode with actual observations obtained from NCEP Reanalysis during that period for uncertainty bounds.

Owing to what are known as cascading uncertainties (the percolation of uncertainties from emissions to carbon and on to climate model projections, followed by regional climate and hydrological models all of the way to assessments of regional impacts), even a small gain in the characterization of uncertainty at the earlier stages of the cascade can lead to significant improvements in risk-informed decision-making related to adaptation and mitigation. Stakeholders and end-users require credible assessments of climate change and extremes at local to regional scales for adaptation decisions (e.g., management of natural water and nutritional resources, development of natural hazards preparedness and humanitarian assistance infrastructures, as well as dealing with climate related migrations) and policy negotiations (e.g., emissions regulations and agreements). The ultimate goal may be to reduce uncertainties by developing precise characterizations of the predictive ability of the current generation of climate models, pointing to areas where enhancements to physical understanding or modeling processes may be necessary, and suggesting pathways for climate model improvements in conjunction with statistical developments for uncertainty treatment.

Geographic Variability. Risk-informed decisions require cost-benefit analyses, for which regional-scale climate projections and impacts become important first steps. The geographic variability of climate change and extremes are important in this context. A recently concluded high-profile climate change war game brought these matters to the forefront (24).

In the main body of the paper, we already illustrated the geographic variability between different IPCC SRES scenarios. Here, we expand upon this discussion by examining variability within the same scenario by comparing the three individual A1FI ensemble members; this is an analysis for ensembles at a global scale. Fig. S3 shows the differences between ensemble outputs 2000 (top), 2050 (middle), and 2100 (bottom). The three columns represent the three possible pairwise comparisons between the ensemble members. First and foremost, we observe that significant spatial variability is indeed present, ranging from $-2.4\,^{\circ}$ to 1.6 °C across all pairings. In addition, note that magnitude and direction, as well as the geographic locality, of observed differences may vary over time. For instance, the third pairing has a large negative difference over much of Europe and northern Asia in 2000 (top right), but by 2100, this pattern has nearly inverted (bottom right), as the majority of the area now shows a moderate positive difference. These time-variant differences of up to 2.4 °C between ensemble members further increase the uncertainty associated with model projections, which should be taken into consideration in particular for regional- or local-scale analyses of future climate change and/or impacts.

Intensity, Duration, and Frequency of Heat Waves. We consider the severity of heat waves based on the A1FI scenario, and Fig. 4 illustrates their current intensity as well as projections including upper and lower bounds for 2050 and 2100. However, in characterizing heat waves, we are often interested in the trifecta of intensity, duration, and frequency for such events. Here we compare heat waves from CCSM 3.0 model outputs to observations and examine the latter two components, i.e., duration and frequency, omitted from the body text for space reasons.

Fig. S4 shows heat waves for observations from 2000–2007 (black) and for A1FI from 2000–2100 (red). The solid line denotes the actual model output, while the dotted line indicates the bias-corrected values with three-sigma error bounds. We note that even the worst-case A1FI scenario under-predicts intensity for the current decade by 0.65 °C on average, with an increase of 4.36 °C by the end of the 21st century. Hence, heat waves may become more intense than previously thought, and impact studies should give this possible scenario, once believed "too extreme," more serious consideration.

For duration and frequency, we adopt a different definition of a heat wave, namely one based on the probability of occurrence. More specifically, we select as threshold the 95th percentile of nighttime minima over the period 2000–2007, and any night exceeding this threshold is considered a heat wave. In Fig. S5, the left column corresponds to average duration (number of days) and the right column to the frequency (number of events per year). The top row represents the reanalysis data for 2000–2007, the next row the A1FI-forced model outputs for the same period, and the bottom two rows the model outputs for 2050 and 2100, respectively. First, we find that there is a fair amount of agreement between the observed and model-based values, both in terms of magnitude and spatial locality. Similar to intensity, we also observe significant increases in both duration and frequency of heat events in 2050, and even more so by 2100. It is important to point out that the decrease in frequency for 2100 (bottom right) along the equator does not represent a true reduction in heat events, but rather is a visual artifact caused by a single heat event lasting the entire year; that is, the nighttime low temperatures in 2100 will exceed the 95th percentile of present levels every single night. Combined with our analysis of intensity in the main body, these insights paint a grim picture for the end of the century if emissions continue to evolve along the trajectory defined by the A1FI scenario.

Computational Issues. Issues of computational complexity arise in a number of different areas related to this work. First, there is the immense complexity of the climate models themselves, which can strain even the most advanced supercomputing resources. In the case of CCSM 3.0, model runs proceed at ≈4.5 years of simulation per day of wall clock time running on 192 processors*—this is the major reason why a Monte Carlo approach using hundreds or thousands of model runs is simply not feasible. And while these numbers are already staggering, there is a continuous push to further reduce both spatial and temporal resolutions, but also to include additional processes that are currently being omitted to keep the computation tractable. In the case of CCSM 3.0, experiments are already being run with output at 6-h intervals, and improvements in the spatial resolution are only a matter of time.

Likewise, there are computational issues associated with the analysis of the data, both model outputs as well as observations. For example, merely calculating the heat waves from 100 years of daily A1FI outputs (20 GB of data) took 60 CPU h on a high-performance compute cluster. If we were to expand this type of analysis to the approximately 100 CCSM 3.0 output variables, it would require the equivalent of nearly 1 year of CPU time—not to mention the fact that most variables are output at each of 26 vertical layers. Moreover, we may be interested in more complex analytic techniques, e.g., the application of extreme value theory instead of an exceedance-over-threshold approach, which would increase the computational complexity of the problem manifold. For these reasons, consideration must

be given to computational requirements and constraints when performing any type of analysis on climate data.

Assumptions and Limitations. In this section, we state all known assumptions in our analysis and limitations of the available data and/or methods.

We assume a Gaussian distribution of differences between model outputs and observations and test the validity of this assumption (shown in *SI Text*). The temperature differences appear to follow the Gaussian quite closely, which is not unexpected. However, this assumption may not hold for other variables like precipitation.

Because bias and variance are stationary in hindcasts (shown in this *SI Text* under *Statistical Methods*), we assume the same will be true for projections as well. This assumption appears reasonable in this context and is not to be confused with the assumption about projection lead times below.

We implicitly assume the uncertainty remains stationary even as a function of projection lead time. While we would expect an increase, this is difficult to quantify as models evolve over time. The assumption is a strong one. However, this does not take away from our analysis of most likely values or our assertions about large uncertainty and geographic variability. If nonstationarity with projection lead times is considered, uncertainty and variability are expected to increase, lending further support to our assertion.

We assume the statistical methods used to compute confidence intervals are valid despite a small number of data points (8) in the sample, namely yearly differences for 2000–2007. This is a strong assumption in general, but given our tests for Gaussian and stationarity, probably not too unreasonable. This may be especially true when global averages are considered. One alternative would be to compute the intervals by comparing model hindcasts and observations. However, this alternative approach has shortcomings as explained in the main paper. First, the models implicitly take into account past observations, and hence a comparison with hindcast may tend to underestimate the uncertainty in projections. Second, hindcasts do not consider the uncertainty caused by the emission scenarios. On the other hand, comparing the most recent (in this case 2000–2007) model outputs to corresponding observations may be considered as close to online validation as may be achievable. Thus, the definitions of IPCC SRES scenarios have remained the same during this period, and the improvements in CCSM3 model specifically to account for observations from 2000–2007 have been minor if any.

Grid-based estimates are only as precise as the model ($1.4° \times 1.4°$ grid), but are interpolated for visualization so that overall spatial trends are more easily discernible. Geospatial visualization of climate data are an important research area and may need to cater to multiple stakeholders like the modeling science community and the consequence analysis community, as well as to policy and decision-makers.

NCEP Reanalysis data are taken as a proxy for observations, even though we are cognizant that these data are not actual ground measurements, but the product of a model applied to observed data from a variety of sources. This assumption may be reasonable for temperature but needs to be tested for variables like precipitation.

Our analysis relies on global averages in certain situations and grid-based analysis in others. However, we use the outputs from general circulation models (GCM) only. Future work needs to examine continental to regional scale analysis, as well as outputs of either regional climate models (RCM) or higher-resolution GCM, when available.

We only use simulations from a single GCM here (namely, CCSM 3.0) but a multimodel ensemble approach is certainly possible and could be used in future research. We conjecture that similar mean trends of global temperature would hold when other models are included, but uncertainty and geographic variability would be further increased.

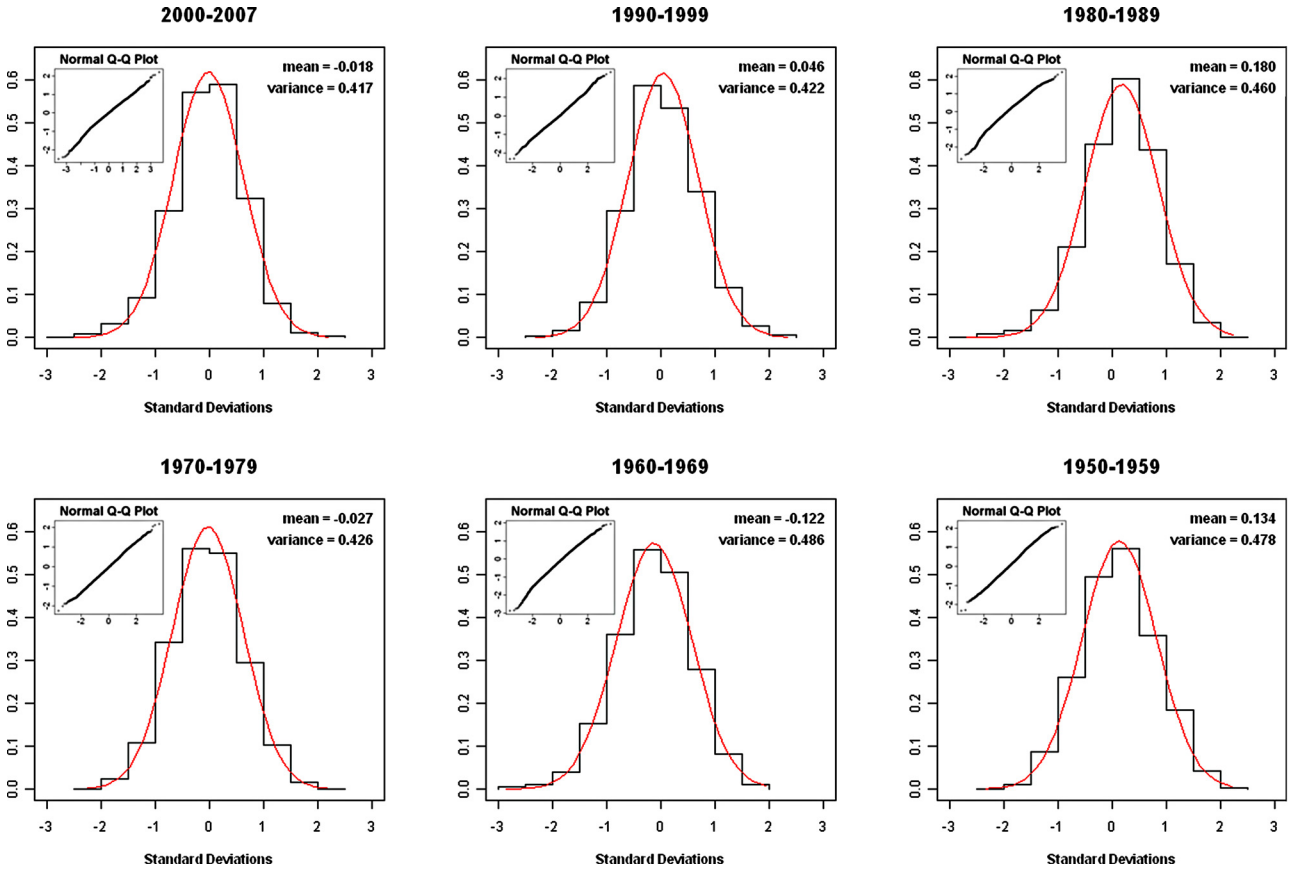

Fig. S1. Histograms of the model-reanalysis bias validate the normality assumption; best-fit curves are shown in red and Q-Q plots as insets.

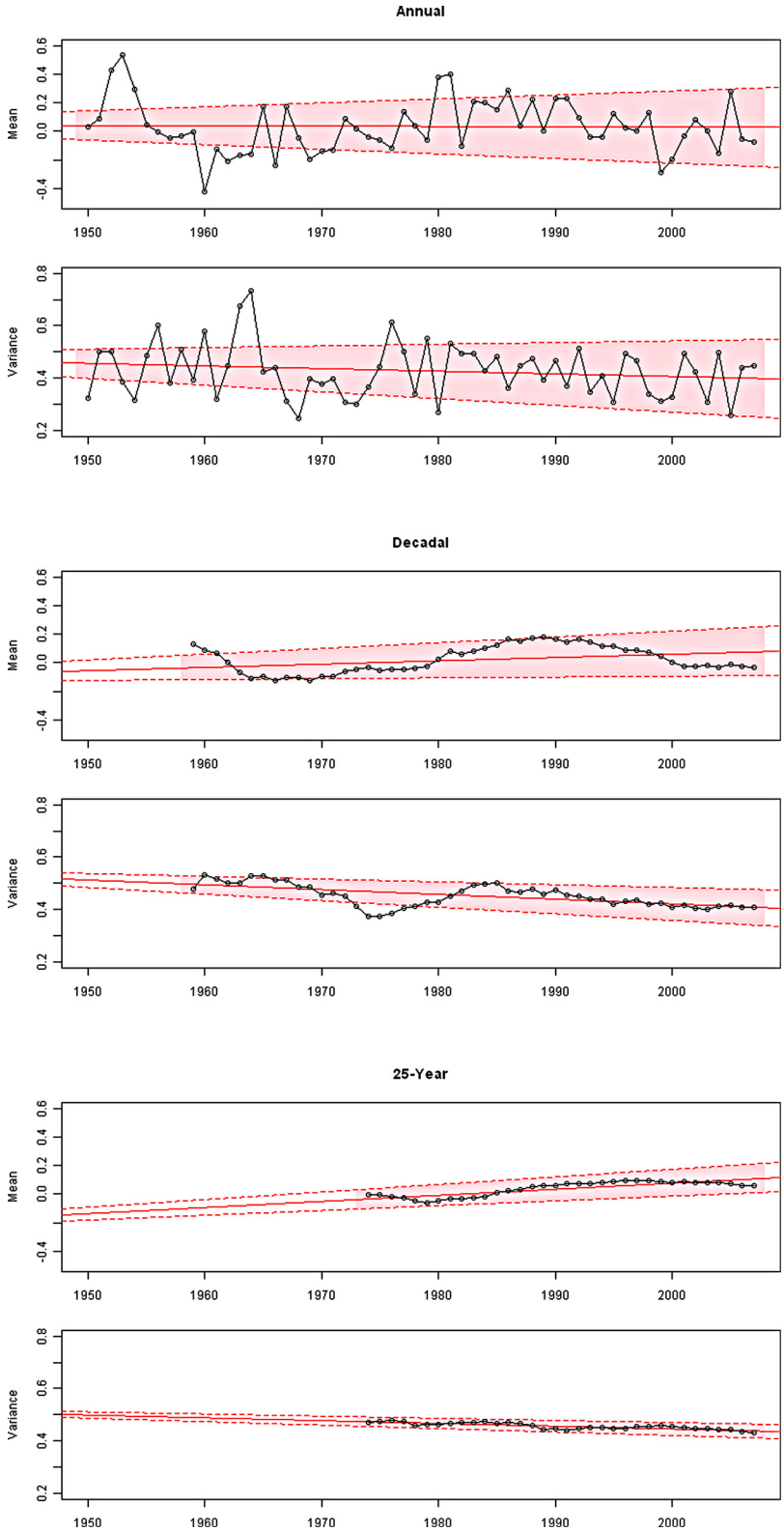

Fig. S2. Analysis of mean and variance of the differences in reanalysis and model simulations computed using 1950–2007 data.

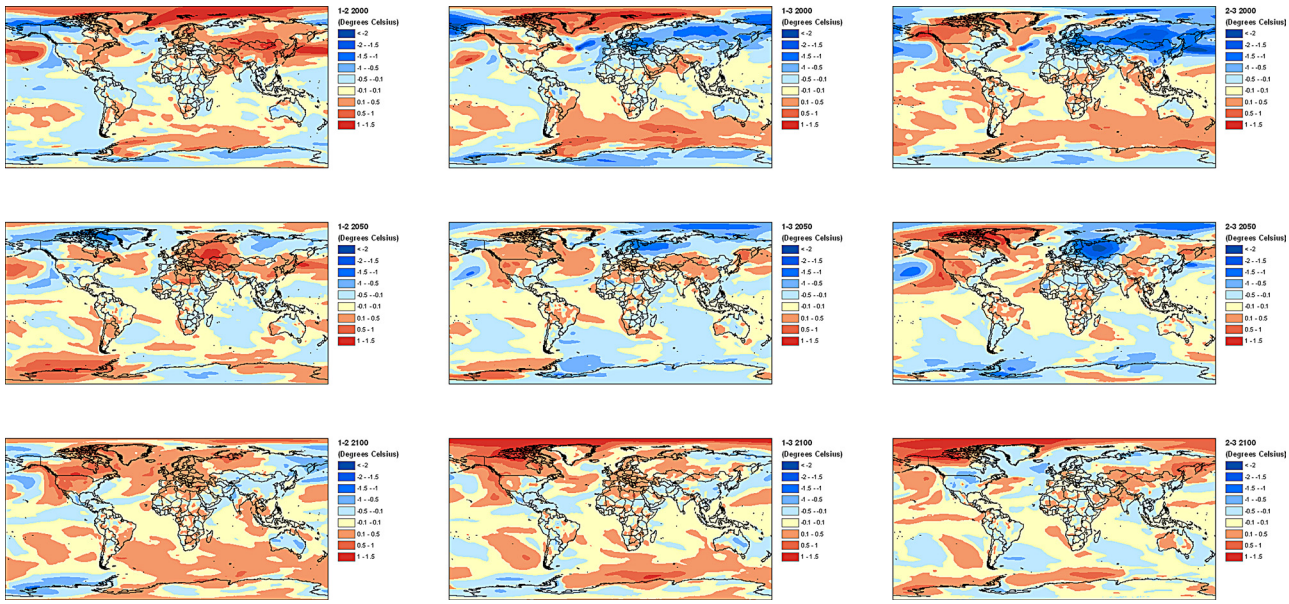

Fig. S3. Geographic variability between three ensembles of the A1FI scenario for 2000, 2050, and 2100.

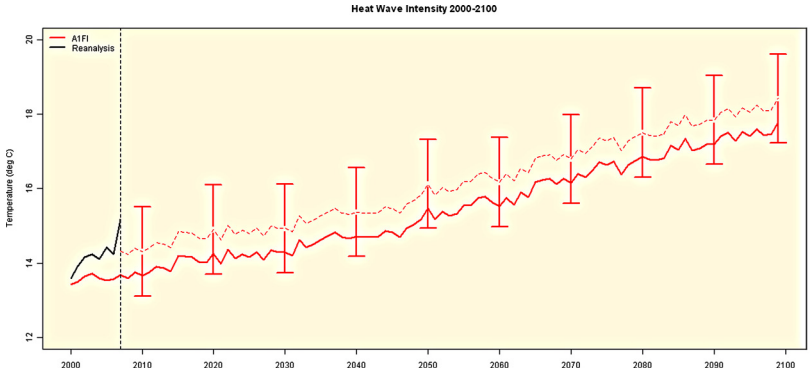

Fig. S4. Globally-average heat wave (°C) projections from CCSM 3.0, based on A1FI, along with error bars. The bias and standard deviations are calculated for each projection by comparing NCEP Reanalysis data with model outputs in 2000–2007, which forms the basis in the generation of the error bars for 2010 to 2100.

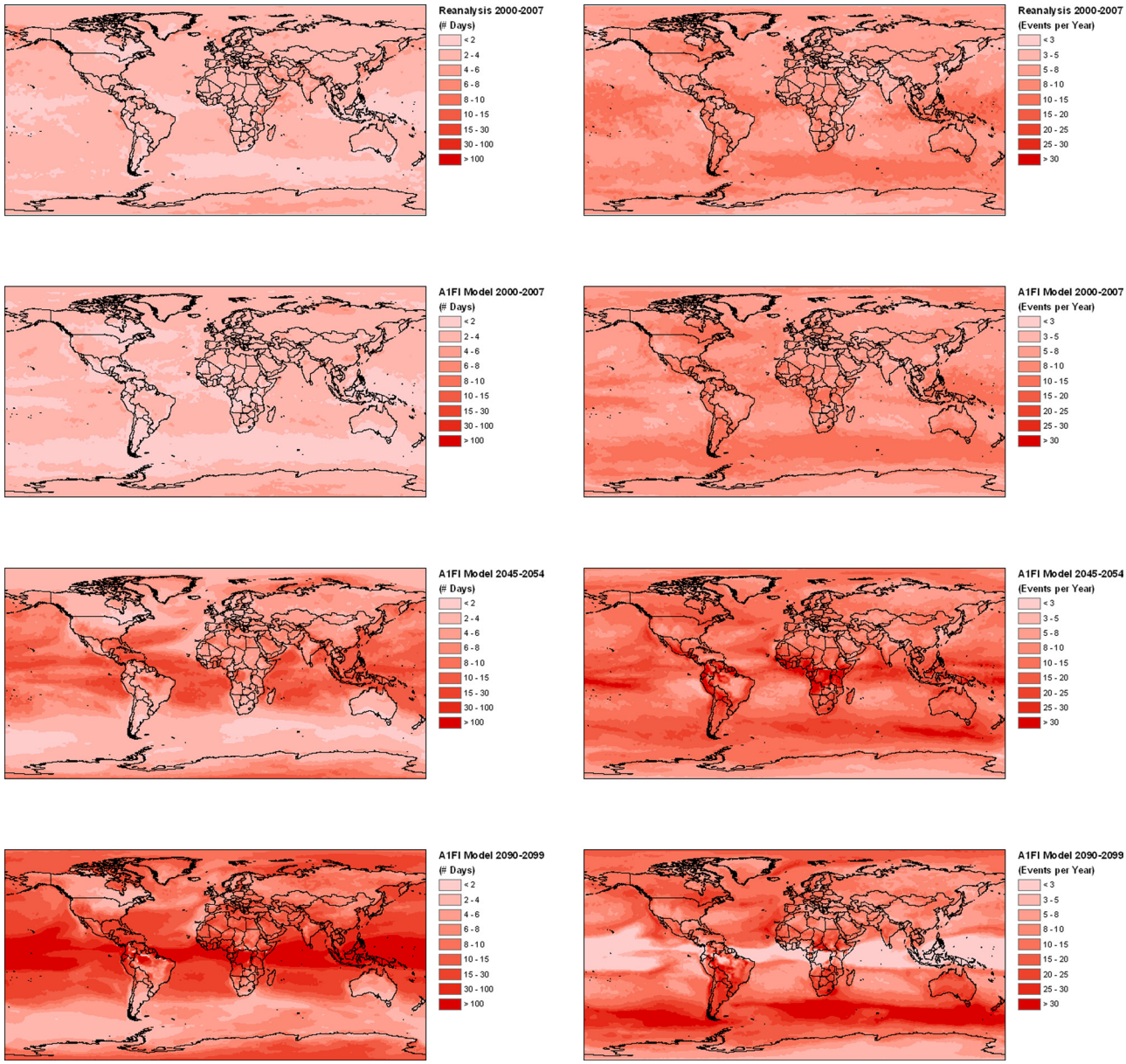

Fig. S5. Duration (*Left*) and frequency (*Right*) of heat waves from NCEP Reanalysis (top) and A1FI for 2000, 2050, and 2100.

Table S1. Pairwise *t* test results between the A1FI, A2, and B1 scenarios at 95% confidence ($\alpha = 0.05$)

		2010	2020	2030	2040	2050	2060	2070	2080	2090	2100
A1FI-A2	*t*	0.385	0.295	0.059	0.599	1.034	1.781	2.018	1.851	2.041	1.371
	95%	No	No	No	No	No	Yes	Yes	Yes	Yes	No
A1FI-B1	*t*	0.485	0.900	0.840	2.978	4.646	6.397	8.575	10.970	13.303	8.640
	95%	No	No	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes
A2-B1	*t*	0.182	0.393	0.945	1.952	2.873	3.334	5.111	7.807	9.821	12.667
	95%	No	No	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Table S2. *f* test results for the A2 and B1 scenarios comparing uncertainty bounds to the ensemble bounds at 95% confidence ($\alpha = 0.05$)

		2010	2020	2030	2040	2050	2060	2070	2080	2090	2100
A2	*f*	30.939	7.356	2.939	7.893	3.863	16.494	11.717	212.74	6.423	9.686
	95%	No	No	No	No	No	No	No	No	No	No
B1	*f*	46.012	5.932	5.975	20.988	28.977	23.393	37.963	9.939	8.870	5.106
	95%	Yes	Yes	No	Yes	No	Yes	Yes	Yes	Yes	Yes